PhD position in CEREA/LMD

Machine learning, data assimilation and dynamical systems

Context

Data assimilation aims at estimating and forecasting a geophysical system by combining in a mathematically optimal way a high-dimensional model and a large observation dataset of that system. This is especially useful for chaotic systems whose horizon of predictability is intrinsically limited. Data assimilation has been hugely successful in improving the skill of weather forecasting for the past 30 years.
More recently, machine learning techniques, and especially deep learning, made impressive breakthroughs in image and speech recognition. These techniques are spreading to many fields in sciences.
Data assimilation and machine learning share common goals and part of their mathematical foundations. The general objective of this thesis is to look at the potential of machine learning techniques for data assimilation and for modelling chaotic geophysical fluids such as the atmosphere.

Project

In image and speech recognition, recurrent neural networks were shown to be potentially remarkable substitutes for dynamical models [Lecun et al. 2015]. It is natural to wonder if machine learning techniques could also apply to high-dimensional chaotic geophysical models. Specifically, neural networks (NNs, [Goodfellow et al., 2014]) could be a surrogate for a whole model or for a part of it, such as the representation of subgrid scale processes (e.g., Gentine et al 2018). In order to determine the coefficients of a NN, one would need to massively observe the geophysical model to be emulated. However, in practice, these observations would be sparse and noisy, potentially much more than in image recognition applications.

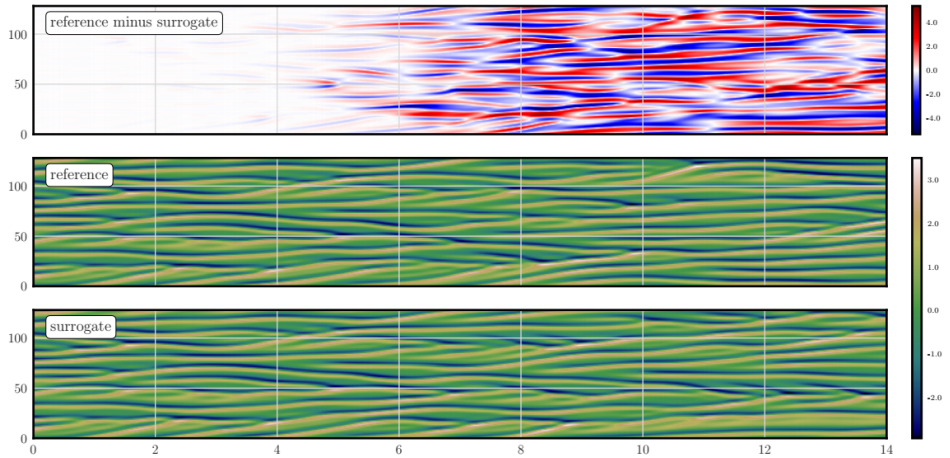

Figure: Comparing the forecasts of the chaotic Kuramoto-Sivashinki model and of a surrogate model obtained from the observation of the true model [Bocquet et al., 2019]. The y-axis spans the 128 variables of the discretised model and the x-axis shows time.

In the wake of very preliminary results [Bocquet et al., 2019 and references therein], the first objective of this PhD is to test if we can combine data assimilation and machine learning techniques to learn the dynamics of a chaotic model. This main objective is declined into subquestions, spanning from methodological to more fundamental:
- How difficult is the learning step? Can we speed it up?

- What are the most relevant representations for the surrogate model? Which architecture should we design for the NN? What are the most relevant variables to work with?
- What are the properties of the surrogate model as a dynamical system? Does it have good shadowing or forecasting skills on the full system? Does its asymptotic behaviour match that of the original model?

These questions will be tackled theoretically and numerically with increasingly complex low-order and intermediate chaotic models such as the Lorenz models, a basic QG model and a more advanced one representing realistic variability modes of the atmosphere [D'Andrea & Vautard, 2001]. The surrogate models obtained from the combined used of data assimilation and machine learning could be compared to the reduced models obtained from the true model using more classical techniques.

A second objective of this PhD is to determine how useful can machine learning techniques be to improve data assimilation techniques, when we already know the model to a large extent. We shall in particular focus on the *ensemble Kalman filter* [Evensen, 2009], a very successful and popular data assimilation technique but which requires fine tunings that could be addressed by machine learning techniques.

Bibliography:

- Bocquet, M.; Brajard, J.; Carrassi, A. & Bertino, L. (2019), 'Data assimilation as a deep learning tool to infer ODE representations of dynamical models', *Nonlin. Processes Geophys. Discuss.* **2019**, 1-29.
- D'Andrea, F. & Vautard, R. (2001), 'Extratropical low-frequency variability as a low-dimensional problem I: A simplified model', *Q. J. R. Meteorol. Soc.* **127**, 1357-1374.
- Evensen, G. (2009), *Data Assimilation: The Ensemble Kalman Filter*, Springer-Verlag Berlin Heildelberg.
- Gentine, P.; Pritchard, M.; Rasp, S.; Reinaudi, G. & Yacalis, G. (2018), 'Could Machine Learning Break the Convection Parameterization Deadlock?', *Geophys. Res. Lett.* **45**, 5742-5751.
- Goodfellow, I.; Bengio, Y. & Courville, A. (2016), *Deep learning*, The MIT Press, Cambridge Massachusetts, London England.
- LeCun, Y.; Bengio, Y. & Hinton, G. (2015), 'Deep learning', *Nature* **521**, 436.

Location:
The candidate will work at École des Ponts ParisTech (Champs-sur-Marne, RER A Noisy-Champs) and École Normale Supérieure (Paris, RER B Luxembourg).

Collaborations:
The PhD will be co-supervised by Marc Bocquet in CEREA, a joint laboratory of École des Ponts ParisTech and EdF R&D and Fabio D'Andrea in LMD (Laboratoire de Météorologie Dynamique) at École Normale Supérieure.

Key words:
Data assimilation, machine learning, deep learning, geofluids, dynamical systems, dynamical systems, chaos.

Duration:
3 years, PhD start: fall of 2019.

Skills and profile: The PhD candiadate must have a master degree in either fluid mechanics, fundamental geosciences, computational physics or applied mathematics. Moreover, the candidate should be comfortable with programming languages such as Python, Fortran and/or C/C++. Some knowledge of deep learning tools such as TensorFlow, Keras, Pytorch would be appreciated.

Contacts:

Marc Bocquet (marc.bocquet@enpc.fr) and Fabio D'Andrea (fabio.dandrea@lmd.ens.fr)